

Rapport Projet 4

Sylvain VASSEUR



16 avril 2025

Projet personnel à but formateur et exploratoire

Table des matières

[I. Résumé exécutif 1](#_Toc195875952)

[II. Introduction 1](#_Toc195875953)

[III. Méthodologie 1](#_Toc195875954)

[IV. Résultats 1](#_Toc195875955)

[V. Analyse approfondie 1](#_Toc195875956)

[VI. Recommandations 1](#_Toc195875957)

[VII. Conclusion 1](#_Toc195875958)

# Résumé exécutif

* **Contenu**: Contexte, objectif, méthodes, résultats clés, recommandations principales
* **Bonnes pratiques**: Limiter à 1 page, être percutant, centrer sur la valeur business

# Introduction

* **Contenu**: Contexte business, objectifs précis, questions de recherche

Nous souhaitons analyser les écarts de performance entre ses départements.

* **Bonnes pratiques**: Formuler clairement les questions auxquelles l'analyse répond

Contexte  
Dans un souci d’amélioration continue, une entreprise a exprimé le besoin d’évaluer les écarts de performance entre ses différents départements. Cette initiative vise à mieux comprendre les facteurs internes qui influencent la productivité et la satisfaction des collaborateurs.

Objectifs  
L’objectif principal de cette analyse est d’identifier les différences de performance entre départements, à partir de variables à la fois qualitatives (comme le type de département ou le niveau de satisfaction) et quantitatives (comme les scores de performance ou les heures travaillées). Il s’agit également de détecter d’éventuelles corrélations ou tendances permettant d’orienter des décisions de gestion ou d’optimisation des ressources.

Questions de recherche

* Existe-t-il des écarts de performance significatifs entre les départements ?
* Quels facteurs (satisfaction des employés, heures travaillées) expliquent les écarts de performance observés entre les départements ?
* Quelles relations spécifiques peut-on identifier entre les variables qualitatives (départements, satisfaction) et quantitatives (scores de performance, heures travaillées) ?
* Peut-on identifier des profils de départements plus performants ou moins performants que la moyenne ?

# Méthodologie

* **Contenu**: Sources de données, préparation, méthodes d'analyse, limites
* **Bonnes pratiques**:
  + Être concis mais précis
  + Expliquer les choix méthodologiques
  + Mentionner les tests statistiques utilisés
  + Décrire les transformations de données

##### Analyse des départements :

**Méthodes :** Statistiques descriptives comparatives, tests d'hypothèses sur les différences (ANOVA, tests t), analyse de distribution des performances

**Objectif :** Identifier et quantifier les écarts de performance entre les départements, en répondant à la question : *"Existe-t-il des écarts significatifs de performance entre les départements ?"*

##### Analyse des facteurs explicatifs :

**Méthodes :** Analyse de corrélation bivariée et multivariée, modélisation des relations entre variables, analyse de régression

**Objectif :** Déterminer l'influence de la satisfaction des employés et des heures travaillées sur la performance, pour répondre à : *"Quels facteurs expliquent les écarts de performance observés entre les départements ?"*

##### Identification des profils de départements :

**Méthodes :** Techniques de segmentation, analyse comparative des groupes de performance, caractérisation multidimensionnelle des profils

**Objectif :** Caractériser les départements surperformants et sous-performants et identifier leurs attributs distinctifs, répondant à : *"Peut-on identifier des profils de départements plus performants ou moins performants que la moyenne ?"*

##### **Synthèse et recommandations :**

**Méthodes :** Analyse croisée des résultats, hiérarchisation des facteurs d'influence, modélisation d'impact des interventions potentielles

**Objectif :** Formuler des recommandations concrètes et hiérarchisées pour réduire les écarts de performance entre départements.

# Résultats

1. Analyse des départements
   1. Analyse des statistiques descriptives

Cette section présente les résultats de l'analyse quantitative des écarts de performance entre départements. Les données ont été analysées à l'aide de statistiques descriptives et de tests d'inférence statistique conformément à la méthodologie décrite précédemment.

Dans un premier temps, les indicateurs de tendance centrale et de dispersion sont examinés pour chaque département. Le Tableau 1 synthétise ces statistiques fondamentales, permettant une première comparaison factuelle des performances observées.

[TABLEAU 1] Tableau stats descriptives

Une remarquable **homogénéité** ressort des données : les moyennes varient dans un intervalle restreint, de 2,98 à 3,02, soit une amplitude de 0,04 points. De même, les écarts-types présentent une variabilité minimale (1,41 à 1,42). Les valeurs minimales et maximales sont identiques pour toutes les catégories, et le nombre d'observations par catégorie s'élève à environ 11 000 unités.

Ces résultats indiquent une **distribution** des performances **très uniforme** entre les neuf catégories, sans écart notable entre elles.

* 1. Tests d’hypothèses sur les différences

Afin d’identifier d’éventuelles disparités entre les départements, un test d’ANOVA est réalisé.  
Ce test permet de détecter si, pour une variable donnée, **au moins deux départements** présentent une **différence significative de moyenne**.

Les variables sélectionnées pour cette analyse sont les suivantes :

* Work\_Hours\_Per\_Week : volume horaire de travail
* Overtime\_Hours : charge ou pression supplémentaire
* Projects\_Handled : nombre de missions gérées
* Performance\_Score : niveau de performance évalué
* Remote\_Work\_Frequency *:* fréquence de télétravail
* Team\_Size *:* effectif moyen par équipe
* Monthly\_Salary *:* niveau de rémunération mensuelle

Ces variables ont été choisies car elles permettent de **caractériser de façon concrète l'organisation du travail dans chaque département**. Un résultat significatif à l’ANOVA indiquerait qu’au moins deux départements présentent des **conditions ou résultats sensiblement différents** sur l’un de ces aspects — ce qui pourrait orienter des décisions de pilotage ou d’équilibrage interne.

[TABLEAU 2] Tableau du test ANOVA

Ces résultats suggèrent qu'il n'y a pas de disparité significative entre les départements en termes de **charge de travail**, **pression supplémentaire**, **nombre de projets gérés**, **performance**, **télétravail**, **taille des équipes**, et **salaires**.

* 1. Analyse de distribution des performances

Bien que certaines variables présentent des p-values proches du seuil de significativité (notamment Projects\_Handled avec p = 0.086), ces différences ne sont pas assez marquées pour être considérées comme statistiquement significatives au seuil classique de 5%. Par conséquent, ces résultats ne justifient pas une distinction notable entre les départements pour les aspects analysés.

**Intéressons-nous à la répartition des scores de performance au sein des départements.** L’analyse statistique suggère que les distributions des scores sont, à quelques nuances près, similaires entre les départements.

[TABLEAU 3] Histogrammes % perf by departemetn

Cette observation est confirmée par les données : la répartition des scores de performance ne montre pas de disparités significatives entre les départements. La proportion d’employés pour chaque niveau de score (de 1 à 5) reste globalement stable d’un secteur à l’autre.

* 1. Synthèse des observations

L’analyse des scores de performance selon les départements révèle une forte similarité entre les groupes. Les statistiques descriptives montrent des moyennes très proches et des écarts types quasi identiques, le tout avec un volume de données similaire entre départements. Les résultats du test ANOVA confirment cette stabilité, aucune des variables testées ne présentant de différence significative entre les groupes. Par ailleurs, la répartition des scores de performance est équilibrée dans l’ensemble des départements, chaque modalité représentant environ 20 % des effectifs. Ces constats ne mettent en évidence **aucune variation notable liée au département à ce stade de l’analyse**. Pour la suite, nous pourrons nous concentrer sur d’éventuels leviers internes aux départements ou sur des variables transverses susceptibles d’apporter un éclairage complémentaire.

1. **Analyse des facteurs explicatifs**
   1. Analyse des variables quantitatives
2. *Corrélations avec le Performance\_Score*

Afin de mieux comprendre quels facteurs peuvent influencer le *Performance\_Score*, nous avons réalisé une analyse de corrélation. L’objectif est d’identifier si certaines variables sont significativement liées à la performance des employés.

[TABLEAU 3] Corrélation de Perforamnce Score

Les résultats sont globalement peu concluants : à l’exception d’une seule variable, toutes les corrélations obtenues sont proches de zéro. Cela signifie qu’aucun lien clair n’émerge entre le score de performance et la majorité des variables analysées. Ces résultats freinent notre capacité à dégager des pistes concrètes d’amélioration.

Seule exception notable : le Monthly\_Salary, qui présente une corrélation de **0.51** avec le Performance\_Score. Cette valeur suggère une relation modérée — autrement dit, plus un employé est rémunéré, plus son score de performance tend à être élevé. Ce lien reste cependant à nuancer : bien qu’une corrélation de 0.51 soit intéressante, elle ne permet pas, à elle seule, d’expliquer pleinement les différences de performance observées.

1. *Régression multivariée*

Afin d’avoir une vision globale de l’impact des variables quantitatives, nous réalisons une régression linéaire multiple, qui permet d’évaluer l’effet combiné de plusieurs variables sur une variable cible – ici, l’**indice de performance**.

**[TABLEAU 3] — Régression multivariée du Performance\_Score**

Nous nous intéressons uniquement aux variables dont la **valeur p** est inférieure à 0.05, ce qui signifie que leur influence est statistiquement significative. Dans notre cas, seules deux variables remplissent ce critère : Monthly\_Salary et Team\_Size.

Le coefficient associé au salaire mensuel peut paraître faible (0.0005), mais il s’applique à une variable aux valeurs élevées. Ainsi, une variation réaliste du salaire entraîne une variation notable du score de performance. Ce type de lecture rappelle que la signification d’un coefficient dépend autant de sa valeur que de l’échelle de la variable concernée.

De son côté, **Team\_Size** affiche un coefficient négatif (-0.0017), ce qui suggère qu’à mesure que la taille de l’équipe augmente, le score de performance individuel tend à diminuer légèrement.

Enfin, le **coefficient de détermination** R² = 0.26 indique que l’ensemble de nos variables explicatives parvient à expliquer **26% de la variance observée** du score de performance. Cela reste une proportion honorable dans un contexte RH, où de nombreux facteurs externes ou qualitatifs (non mesurés ici) peuvent influencer la performance.

1. *Analyse par département*

Nous poursuivons notre analyse en introduisant la dimension des départements. Pour cela, nous réalisons une régression linéaire (OLS) sur le score de performance, en isolant chaque département. Cette approche permet d’identifier les spécificités propres à chaque unité, qui pouvaient être noyées dans l’analyse globale.

Pour chaque département, deux résultats clés sont obtenus :

Le **R²**, qui indique la part de variance du score de performance expliquée par nos variables descriptives.

La liste des **variables significatives**, c’est-à-dire celles dont l’impact sur la performance est statistiquement prouvé (p-value < 0.05).

**[TABLEAU 3] Régression linéaire par département – Score de performance**

Les résultats montrent que le R² reste relativement stable, autour de 0.26, ce qui suggère une cohérence globale : dans chaque département, nos variables expliquent environ 26 % du score de performance. En revanche, on observe l’apparition de nouvelles variables explicatives selon les départements. Par exemple, *Sick\_Days* devient significatif dans *Customer Support* et *Legal*, alors qu’il ne l’était pas au niveau global.

Cela s’explique par le fait qu’en segmentant par département, on réduit l’effet de dilution dû à la diversité des situations. Certaines variables, peu visibles à l’échelle globale, gagnent en importance une fois replacées dans leur contexte métier. Cela confirme la pertinence d’une analyse segmentée pour mieux comprendre les leviers de performance propres à chaque service.

Enfin, il est intéressant de noter que le nombre de variables significatives n’a pas de lien direct avec le R² obtenu. Un département avec une seule variable explicative peut avoir un R² comparable à un autre en ayant deux. Cela signifie que certaines variables, bien que significatives, n’apportent qu’une faible contribution à la variance expliquée.

* 1. Analyse des variables qualitatives

1. *Répartition des variables qualitatives*

Nous poursuivons l’analyse en nous intéressant à la répartition des variables qualitatives au sein des départements. Cette étape permet de mieux comprendre la variation des caractéristiques et d’identifier d’éventuelles tendances ou disparités.

**[TABLEAU 3] — Distribution des variables qualitatives par département**

* **Genre :** La répartition par genre est équitable entre les départements, avec un pourcentage similaire d’hommes et de femmes. La catégorie "Other" reste faible, ce qui est un résultat attendu compte tenu de la faible représentation de cette catégorie dans l’ensemble.
* **Niveau d'Étude :** Le niveau d’étude "Bachelor" est majoritaire, représentant près de 50% des effectifs dans chaque département. La catégorie "High School" enregistre environ 3000 employés par département. La distribution des niveaux d’études présente une homogénéité remarquable entre les départements, quel que soit le niveau d’étude (PhD, Master, Bachelor, ou High School).
* **Postes :** Les intitulés de postes ("Job\_Title") montrent une répartition homogène à travers les départements, sans variation notable.
* **Résignations :** Les données sur les départs volontaires ("Resigned") indiquent une répartition stable, avec environ 1000 départs dans chaque département, sans variation significative entre eux.

1. *Test d’indépendance (chi²) global*

Nous poussons nos recherches et réalisons un test di Chi² afin de déterminer si parmi ces variables certaines sont lié au Performance Score :

**[TABLEAU 3] — Test du Chi² sur les variables qualitatives**

Deux variables ressortent particulièrement de cette analyse : le genre et le niveau d’études. Les p-values associées à ces variables sont significatives, avec une p-value pour le genre inférieur à 0.05, respectant ainsi le seuil de significativité. En revanche, le niveau d’études à une p-value de 0.06, ce qui ne respecte pas strictement le seuil de 0.05, mais reste suffisamment proche pour justifier son inclusion dans l’analyse.

Cela nous permet donc d'identifier le genre comme une variable significative à inclure dans l'évaluation de la performance des départements. Le niveau d’études, bien que légèrement au-dessus du seuil, peut également être pris en compte dans l’interprétation des résultats.

1. *Test d’indépendance par département*

Nous pouvons maintenant aller plus loin en réalisant un test du Chi² pour examiner l'indépendance de ces variables par rapport aux départements. Cela nous donnera une perspective supplémentaire sur la relation entre ces facteurs et la performance des départements.

**[TABLEAU 3] — Test du Chi² sur les variables qualitatives par départements**

L’analyse des relations entre les variables qualitatives (Job\_Title, Gender, Education\_Level, Resigned) et le Performance\_Score, menée séparément par département, indique que **seul le département *HR* présente des relations statistiquement significatives** (p-value < 0.05).

Dans ce département, les variables Job\_Title et Resigned sont liées au score de performance.

Il est à noter que la variable Education\_Level, qui présentait une p-value proche du seuil de significativité (0.06) dans l’analyse globale, **n’a pas confirmé cette tendance lorsqu’elle a été examinée séparément par département**.  
À chaque segmentation, les p-values associées à cette variable sont restées **largement supérieures à 0.05**, ce qui ne permet pas de retenir une relation significative entre le niveau d’éducation et le score de performance dans un cadre départemental. Une situation similaire a été observé pour la variable Gender, dont la p-value sur le test du Chi² global était à 0.04. Même si par départements ses valeurs sont relativement proches du seuil de significativité (0.11 à 0.15) elles restent néanmoins **trop élevées** pour conclure à une relation statistiquement significative.

Ces résultats suggèrent que **ces variables n’ont pas un impact significatif** sur la performance, du moins dans le cadre de cette **analyse segmentée par département**.

**Aucune autre relation significative n’a été observée dans les autres départements.**

Cette analyse met en évidence une **disparité entre départements** quant à la présence de variables qualitatives associées à la performance.

* 1. Résultats d’analyse

L’analyse des variables quantitatives met en évidence **des résultats contrastés**. À l’échelle globale, seule la variable Monthly\_Salary présente une corrélation notable avec le Performance\_Score, avec un coefficient de 0.51, tandis que les autres variables affichent des corrélations proches de zéro. Cette observation est confirmée par la régression multivariée, qui retient également Team\_Size comme variable significative, bien que son effet reste modéré. Lorsqu’on isole les départements, le R² reste stable autour de 0.26, ce qui indique une capacité explicative comparable dans chaque unité. Toutefois, de **nouvelles variables deviennent significatives localement**, comme Sick\_Days dans *Customer Support* et Legal, Years\_at\_Company dans *IT*, ou encore Team\_Size dans *Finance*. Ces résultats révèlent que **certains effets se manifestent uniquement dans des contextes spécifiques**, masqués dans l’analyse globale.

Les variables qualitatives présentent une **distribution relativement homogène** entre départements. Le test du Chi² global révèle une association significative entre le genre et le score de performance, tandis que le niveau d’étude approche le seuil de significativité sans le franchir. En affinant l’analyse **au niveau départemental**, **seules** **deux relations apparaissent significatives**, toutes deux dans le département *RH* : les variables Job\_Title et Resigned y présentent un lien statistique avec le Performance\_Score. Aucune autre association notable n’est détectée dans les autres départements, y compris pour Education\_Level, dont l’effet ne se confirme pas localement.

Dans l’ensemble, les données quantitatives permettent d’identifier quelques facteurs explicatifs, dont certains conservent leur influence à travers les départements, tandis que d’autres **ne deviennent visibles qu’une fois les unités analysées séparément**. Les données qualitatives, plus constantes dans leur répartition, révèlent peu de relations significatives avec la performance, en dehors de cas très localisés. Ces résultats illustrent **les limites des variables observées** pour expliquer la performance individuelle, tout en soulignant l’intérêt d’analyses segmentées.

1. **Identification des profils par départements**

Après avoir déterminé les variables significatives pour nos différents départements, l’étape suivante consiste à identifier des profils types. Pour chaque département, nous calculerons le R² afin de mesurer l’impact des variables explicatives sur la variance du score de performance. Ce calcul nous permettra de quantifier la part de la performance expliquée par les variables sélectionnées, et d’évaluer si certains départements présentent des profils distincts de surperformance ou de sous-performance en fonction de leur capacité explicative.

* 1. Variables significatives par département

À partir des analyses précédentes, nous pouvons dresser un tableau récapitulatif associant à chaque département les variables explicatives du score de performance:

* IT : Monthly\_Salary, Years\_At\_Company
* Finance : Monthly\_Salary, Team\_Size
* Customer Support : Monthly\_Salary, Sick\_Days
* Engineering : Monthly\_Salary
* Marketing : Monthly\_Salary
* HR : Monthly\_Salary, Job\_Title, Resigned
* Operations : Monthly\_Salary
* Sales : Monthly\_Salary
* Legal : Monthly\_Salary, Sick\_Days

Trois groupes se dégagent selon le nombre de variables explicatives identifiées.

**[TABLEAU 3] — Résultats du R² sur les variables explicatives par département**

Dans un premier temps, nous avons cherché à analyser les R² pour chaque département afin d'identifier l'impact des variables explicatives sur le score de performance. Pour obtenir les résultats les plus précis possibles, nous avons appliqué du **one-hot encoding** sur des variables telles que Gender (même si elle ne s'avère pas significative ici) et Job\_Title. Cette méthode a permis d'éviter d'attribuer une importance injustifiée à certaines catégories de ces variables, en leur attribuant des valeurs binaires équivalentes et en traitant chaque catégorie sur un pied d'égalité.

Bien que certaines variables, comme **le salaire** (ce qui était attendu), montrent un **R² modéré**, l'ajout d'autres variables significatives n'a pas permis d'améliorer substantiellement la capacité explicative des modèles. Les résultats obtenus restent globalement faibles, ce qui indique qu'aucune relation marquée entre les variables et la performance n'émerge à ce stade. Par conséquent, il n'est pas possible d'établir des profils distincts entre les départements.

La suite logique, dans le cas où certaines variables auraient montré des relations plus intéressantes, aurait été d'analyser **le recouvrement** et les **VIF** pour évaluer si certains départements pouvaient être mieux expliqués. Toutefois, étant donné que ce n'est pas le cas, il n'y a pas lieu de pousser davantage l'analyse sur ce point, et nous passons donc à la suite.

Même avec les variables censées être explicatives de chaque département, il est impossible d'identifier des profils distincts, ce qui rend **difficile la distinction** entre départements **surperformants** ou **sous-performants**.

* 1. Synthèse desprofils

**Penser à ajouter une partie b ? Faire une synthese plus importante ?**

1. **Synthèse et recommandations**

# Analyse approfondie

### 🎯 IDEE :

Comparer les **départements les plus "riches" en variables explicatives** (avec plusieurs relations significatives) avec ceux qui en ont **peu ou pas**, pour :

* Évaluer **jusqu'où peut aller l’explication de la performance par les variables disponibles**,
* Et construire un **intervalle de "pouvoir explicatif"** du Performance\_Score selon les départements.

Tu dois passer par **un modèle explicatif** (ex. régression linéaire multiple ou logistique).  
Mais pour ça, **il faut convertir tes variables qualitatives en variables numériques**, par exemple via :

* **Contenu**: Interprétation des résultats, relations causales, facteurs explicatifs
* **Bonnes pratiques**:
  + Établir des liens entre différentes observations
  + Approfondir les éléments les plus significatifs
  + Tester et valider les hypothèses
  + Éviter de répéter les résultats, les référencer

# Recommandations

* **Contenu**: Actions concrètes, priorisation, mesures de suivi
* **Bonnes pratiques**:
  + Lier directement aux analyses
  + Quantifier l'impact attendu quand possible
  + Prioriser par impact/effort

# Conclusion

* **Contenu**: Synthèse, limites, perspectives futures
* **Bonnes pratiques**: Suggérer les prochaines pistes d'analyse